Building Quality of Service Networks

Using Non-traditional Methods

by

Richard A. Carlson
Electronics and Computing Technology Division
9700 South Cass Avenue
Argonne National Laboratory
Argonne, IL 60439

The author is employed by Argonne National Laboratory, which is supported by the U.S. Department of Energy under Contract W-31-109-Eng-38.

Building Quality of Service Networks Using Non-traditional Methods

Abstract

The traditional view of network Quality of Service (QoS) is that it can only be obtained by designing network components that support and enforce complex administrative policies that restrict the amount of traffic allowed into the network. Applications must determine their networking needs and then signal this need to the network. The network uses a complex set of criteria (i.e., user identity, available resources, etc.) to determine if the application can use the network. The basic assumption is that strict admission control policies are required to manage a high QoS network.

This paper challenges that basic assumption and claims that high-speed network components are no longer scarce resources. Instead they are becoming cheaper and more prevalent in today's networks (i.e., following Moore's Law). In addition, the network end points (i.e., computers) are becoming active components in the management of today's networks. These changes mean that future networks will support application level QoS guarantees without relying solely on complex admission control policies. Such networks will perform more reliably, and will be cheaper to purchase and maintain than those built with traditional QoS mechanisms.

Background

The commercial success of the Internet has proved the value of questioning the traditional view of network engineering [1]. At the time engineers conceived of the Internet, the traditional view was that packet switched networks would never operate on a global scale. It was said that they lacked the basic centralized command and control system required for global network operations. Today the Internet is growing at 100% per year, while the traditional telephony networks are growing at 8-10% per year. It is expected that the total amount of data traffic carried over the Internet will surpass the total amount of voice traffic carried over the telephony networks in the next few years [2].

This rapid growth is changing the way individuals learn, work, and interact with their government representatives. It is also putting new demands on the Internet. One of these demands is the need to carry real-time audio and video traffic over the network in addition to the bulk data traffic currently handled.

The traditional view holds that for the network to support these new, real-time services, it must be enhanced to provide Quality of Service (QoS) guarantees. Further, these enhancements should take the form of strict admission control functions that limit the amount of traffic the network will carry. This view is based on several assumptions:

- Real-time applications are unable to perform without QoS guarantees. Without these guarantees, the receiver will be unable to recover the original data because of the inherent loss and delay of Internet packets.

- Network core components are expensive and scarce resources that must be administratively controlled to prevent network congestion.

At this point it will be beneficial to review the operation of the traditional Plain Old Telephone System (POTS) network and the Internet.

Telephony

The underlying network infrastructure of the telephone system is fiber-optic cables strung between switching centers, forming point-to-point trunks for aggregate traffic. These trunks use Time Division Multiplexing (TDM) technologies to aggregate traffic from many users onto these shared trunks. The use of TDMs means that the network has a fixed number of slots for user data traffic. If all these slots are filled, then no additional traffic may be carried. A central management and control system exists that explicitly distributes routing information to all the switching centers, allowing calls to be established between two end-stations. Upgrading trunks means buying and installing either expensive TDMs or additional fiber-optic cables, or both, between switching centers.

Each individual end-station (telephone) uses only a small, fixed portion of the aggregate bandwidth, typically 64 Kbps for a digital voice channel. This bandwidth is allocated at the beginning of the call and is held by the user until the call is completed. During this time, no other user may use this channel. Since the trunk bandwidth is fixed, the number of voice channels is also fixed. If all of these channels are used, then no additional calls may be created. The allocation of channels depends on the admission control policies designed into the core telephony switches. These switches assign calls to TDM channels on a first-come, first-served basis. In addition, the end-stations do not participate in the network control functions, and they can be positively identified (the phone number is fixed), enabling easy accounting and billing.

The design of telephony networks assumes that the traffic patterns change slowly over time and that it is possible to construct networks to match this demand. Historical data, projected growth patterns, and a fudge factor are combined to design networks that give a low probability of a call failing. This approach typically results in network links being over-provisioned to assure this low probability of call failure. If a failure does occur, the user is informed that he or she must retry the call at a later time. In times of high demand, such as during a natural disaster, the network will not be able to handle the additional load, and more calls will be rejected than normal.

In summary, the telephony network allocates small, fixed-size blocks of network bandwidth to individual users on a first-come, first-served basis. The core network switches manage the number of calls to insure that only a limited number of simultaneous calls are accepted. Individual users enjoy a high QoS guarantee because of the administrative controls used to limit access to the network at all times.

Internet

The underlying network infrastructure of the Internet is the same fiber-optic cables used for the traditional POTS network. The difference is that these cables are connected to statistical multiplexers called Internet Protocol (IP) [3] routers. Each router examines the incoming datagrams and uses header information to determine how to forward them toward the destination. The routing information needed to perform this forwarding task is calculated locally using information learned from the other directly connected “peer” routers. This distributed command and control system means that there is no fixed path through the network that can be used with any guarantee.

Each user end-station (computer) must be configured with local addressing and routing information before it can communicate with a peer end-station. This makes the end-station an intelligent participant in the management and control of the network.

Applications are not designed to operate at fixed rates. Instead the transmission rate varies over time depending on such factors as local link speed, available buffer space in the local and remote end-stations, and packet loss due to network congestion [4,5]. If packets are lost, the transmission rate rapidly decreases, then slowly increases again when no loss is detected.

The demand for network resources varies rapidly as applications “hunt” for their maximum transmission rate. This variability makes it difficult or impossible to design a network that is the proper size for the average number of connections, as was done for the telephony networks. However, with intelligent use of queuing and packet discard algorithms [6], it is possible to design a network that allows all competing applications a fair share of the bandwidth of a trunk.

In summary, the Internet has no fixed paths between any two end-stations. All the network components (i.e., end-stations and routers) participate in a distributed command and control system to send packets from source to destination. When congestion occurs, the core routers use packet discard mechanisms to communicate with end-stations and modify the amount of traffic being injected into the network. Thus, the concept of “best effort” delivery is used to describe how packets are delivered through the Internet.

Experiments to Increase the Internet’s Quality of Service

The desire to support real-time services has led the Internet Engineering Task Force (IETF) to create the Integrated Services [7] (Int-serv) and Differentiated Services [8] (diff-serv) Working Groups. These working groups are defining ways to introduce QoS guarantees into the IP network. The goal is to manage the network during times of congestion to prevent loss of real-time packets and thus maintain application performance at acceptable user levels. If successful, this strategy will allow the Internet to replace the existing POTS network.

Briefly, the int-serv approach attempts to overlay a circuit-switched network onto the Internet by reserving system resources in the network components (e.g., switches and routers). This overlay network uses the Resource ReServation Protocol (RSVP) [9] to signal application requirement to the network components and reserve network resources in these components. Because these resources are held for a short time, whether they are used or not, only a limited number of reservations may be accepted. This approach has been described by its creators as “scheduled unfairness” and “contracted unfairness.” The major drawbacks to the int-serv approach are:

- Policy issues regarding who can set up resource reservations.
- Identification of users for purposes of accounting, billing, and charging for resources used.
- Problems with cross-domain resource allocation.
- Scaling issues for core network components because of the need to store per-flow state information.

The diff-serv work is an attempt to simplify some of the problems encountered in the int-serv model, primarily the requirement for holding information in core network components on per-flow state. By aggregating traffic at domain boundaries, the scaling issue becomes tractable. In addition, cross-domain issues can be reduced to a simple two-party policy problem, where a simple Service Level Agreement (SLA) can be created to specify which traffic is given preference. By using some bits in the IP header [10], the source domain can mark the packets to receive preference, and the carrier domain can simply forward packets with different probabilities on the basis of these IP header markings. The problems with diff-serv are:

- Policy issues for who can set bits in the DS field..
- Identification of users for purposes of accounting, billing, and charging for resources used.
- Re-mapping of bits in the DS field at domain boundaries.

While int-serv and diff-serv policies will provide applications with QoS guarantees, they do so in different ways. Int-serv policies are based on reserving network resources and maintaining state in core network components. Diff-serv policies are based on maintaining multiple queues in the network components so that, during times of congestion, the priority traffic experiences lower loss. If congestion does not occur, then all packets, both high and low priority, will see the same level of network service, and applications will not notice a difference.

Problems with Implementing Quality of Service

The creation of a global Internet with QoS guarantees poses a major problem for network engineers and Internet Service Providers (ISPs). To understand why, we must look at the distributed nature of the many autonomous groups that make up the Internet. From the small ISPs to major corporations and national governments, the Internet is composed of many different voices with many different needs and capabilities. This situation contrasts starkly with the position of the national telephone companies that created the single-service, QoS-based voice network we use today. What were simple issues for telephony companies are major problems for ISPs, ones that must be solved before any QoS Internet can be created. The following are some of the major issues.

- Authorization. The traditional QoS model for the Internet calls for multiple levels of service to be supported over the network. Each level will be priced at a different rate, and users will be allowed to pay for improved service. If this model is to succeed, then some method must be created to allow users to authenticate themselves to the network to prove that they are authorized to pay for the service they use. If all the traffic stays within a single administrative domain, then a simple authorization scheme may be used. However, in the general case, many administrative domains will be crossed, and each domain will need to verify the users' authority. This means that a complex "web of trust" must be created to support this network service.

- Non-repudiation. The traditional telco network relies on the centralized configuration and control of the basic network services, including assignment of station addresses. In contrast, the Internet uses a distributed configuration and control system that allows users to see and change their own station address. This has led to the well-known problems of Smurf and Denial-of-Service attacks in which intentionally misconfigured computers inject traffic into the Internet. If these attacks were launched into a QoS network, then not only would network resources be used, but a bill would be sent to the legitimate owner of the address. Without built-in non-repudiation methods, it will be impossible for the user or carrier to prove or disprove the accuracy of a bill.

- Traffic generation. The traditional QoS model uses a well-defined notion of how much traffic is going to be sent and received over the link. This model allows the source and destination to negotiate the transfer rate before data is sent, and it allows the user to balance the cost vs speed of a data transfer. This model works well if the user knows how much data is to be transferred. In today's point-and-click world, this knowledge is severely restricted. As an example, how could you find out how much data will be sent to you when you click on a web URL? This lack of knowledge will make it very difficult for users to control costs.

- Cost issues. The traditional QoS model uses a switched circuit from source to destination. This circuit allows the network to identify who is to be billed for the service. On the Internet, no such circuit is established, so network routers must maintain state to allow billing to take place. In addition, the Transmission Control Protocol (TCP) used today is a bi-directional service overlaid on the uni-directional IP network. To achieve application-level QoS, the returning acknowledgements must be given the same level of service as the original data packets. If acks are delayed, then TCP may fall into its congestion control algorithm and reduce the application performance below the user's expectations.

These problems, and others yet to be identified, will need to be solved before networks that support strict QoS guarantees can be created. The cost of identifying and solving all these problems must be considered when comparing traditional QoS networks with the non-traditional method described below.

Quality of Service from a Different Perspective

The underlying assumption of Internet QoS systems proposed to date is that congestion occurs in core network components. This congestion leads to packet delay and/or loss and affects all applications in random and unpredictable ways. The traditional QoS view is that to support real-time services, the network must be configured to give “better than best effort” service to some packets at the expense of others.

The question most often asked is how to create and enforce the admission control policies that allow QoS networks to operate. This paper asserts that this is the wrong question to ask when considering Internet QoS networks. The real question is what changes are occurring that will allow real-time applications to operate over the Internet without strict admission control policies being deployed. An additional question is, are these changes continuing at a rate that keeps up with application demands.

The network research into optical network components, high-speed switch and router construction, and high-performance end-station interfaces are three areas that show it is economically possible to build high QoS networks without requiring these traditional admission control policies. It shows that terabit and petabit networks will be available in the near future [11]. It also shows that network speed is doubling every 6 months far exceeding the gains predicted by Moore's law [12]. By coupling intelligent end-stations to this high-speed network, applications will obtain high QoS guarantees without having to resort to traditional QoS methods.

The rest of this paper describes the ongoing research being performed in these three areas.

Transmission Capabilities

The current generation of network links is based on the telephony model of Synchronous Optical Network (SONET) links with multiple 64 Kbps channels. Each link is viewed, and priced, as a separate bundle of these 64 Kbps voice channels. Changes and upgrades to this network are expensive for the following reasons.

- Higher speeds mean more gear. A higher trunk speed means that more channels can be aggregated onto that trunk. This change requires that more switching gear be added to the switching center to handle the additional individual channels. In addition, the regenerators (described below) must also be upgraded or replaced to support the higher speed trunk links.

- Electro-optical regenerators are required at various points in the link to recover from signal propagation losses. These regenerators convert the optical signals into electrical signals, remove any noise, convert them back to optical signals, and transmit them toward the destination. This electro-optical conversion is very expensive and must be done quite often to maintain acceptable signal levels at the receiver.

These two problems are being addressed in ways that will dramatically change how future network links will be built.

The first change is the introduction of Wavelength Division Multiplexing (WDM) to increase the aggregate bandwidth of fibers already in the ground. State-of-the-art commercial SONET systems run at Optical Carrier 192 (OC-192) rates (10 Gbps). Research systems are exploring OC-768 rates (40 Gbps), but the electrical-to-optical conversion presents major hurdles that need to be addressed before these systems can move into commercial use. Increasing transmission speeds to OC-768, and beyond, is also hampered by problems with high-speed electrical switching.

One way around this problem is to increase the aggregate network bandwidth by using multiple, lower-speed links running in parallel. The traditional method was to run parallel optical fibers, with each fiber carrying a single optical signal. Because of recent advances in solid state lasers and photodetectors, multiple fibers are no longer necessary. Instead, these lasers and photodetectors can be tuned to specific frequencies that remain stable over a long period. These multiple frequencies can then be combined and transmitted over a single fiber without interference.

These WDM systems have been around for many years, but the recent advances in laser stability have led to the creation of Dense WDM (DWDM) systems with many channels (32 to 128). Each of these channels can be modulated with its own analog or digital signal at up to OC-192 rates. It is easy to see that using this multi-channel approach makes terabit networks feasible (i.e., 128 x 10 Gbps = 1.28 Tbps). And this rate will only go up as more channels are added. The theoretical maximum limit for such DWDM systems has not yet been determined.

The second change is the development of optical Erbium Doped Fiber Amplifiers (EDFAs), which replace existing, and expensive, electro-optical regenerators. As optical signals travel down a fiber, they are attenuated and need to be amplified for long-distance communications. The traditional method of amplification is to regenerate the signals by using electro-optical converters at various points in the transmission path. In DWDM systems, regeneration is a very expensive operation, as each wavelength must be recovered and regenerated separately. Increasing the number of wavelengths requires upgrading or replacing all the regenerators in the link.

The advent of the EDFA means that optical signals can be directly amplified, removing the need for electro-optical conversion. This advance greatly reduces the cost of building and maintaining DWDM networks. Adding a new channel is simply a matter of upgrading the two end points of the link. In addition, the transmission rate or type (analog or digital) can be easily changed at the end points. In a traditional system, such changes would require upgrading each regenerator. By using these all-optical EDFAs, fiber cable runs of hundreds of miles can be created at prices far below previous rates.

By combining DWDMs and EDFAs, it is possible to build high-speed wide area networks at a fraction of the cost of traditional telephony networks [13]. In addition, costs are further reduced by replacing multiple, low-speed TDMs with a single, advanced IP router.

Routers

The current IP routers have much improved packet-forwarding capabilities: they can move hundreds to thousands of packets per second. Part of this increase comes from merging layer 2 switch architectures with the traditional layer 3 router architectures. This improvement occurred, in part, because designers realized that both the IP header format and its location in the layer 2 protocol frame are well known. Thus, it is possible to build hardware that easily accesses this part of the received packet and makes a rapid decision about forwarding. The router no longer has to read and write the packet in a software forwarding system.

The next generation of IP routers [14] under development today will operate at speeds in the range of millions to billions of packets per second. To reach these speeds, several improvements are being developed, including multiple packet-forwarding engines, multiple copies of the routing tables, high-speed internal switch architectures, and intelligent queuing and forwarding algorithms. The new routers differ from traditional routers in four important ways.

1. Traditional routers use a single central processor to manage all the routing functions. These functions include route calculations, maintenance of routing tables, and slow-path packet-forwarding functions. Simple, on-board caches allow fast-path packet forwarding to occur once a data flow has been detected. The next-generation routers will use multiple full-function forwarding engines—either Application Specific Integrated Circuits (ASICs) or high-performance processors—to maintain high packet-forwarding rates. Routing table calculations and maintenance functions will be performed by a separate dedicated processor.

2. Traditional routers have multiple line cards interconnected by a shared bus architecture. Next-generation routers will replace this shared bus with a high-speed internal switch or shared memory architecture. In these routers, multiple, parallel data paths will exist, improving packet-forwarding performance.

3. Traditional routers use a single routing table that is supplemented by small caches installed on line cards to improve packet-forwarding performance. The advent of low-cost, high-density memory subsystems means that the small caches can be replaced with duplicates of the centralized routing table. Coupled with this change are new lookup algorithms that can efficiently scan these large routing tables to maintain high forwarding rates.

4. Traditional routers use a simple FIFO queuing mechanism at each output port to forward packets. They also have simple tail-drop policies for discarding packets if queues overflow. Next-generation routers will have multiple queues or intelligent packet selection algorithms to determine the packet-forwarding behavior. In addition, intelligent discard (i.e., RED or RIO) and discard notification (i.e., ECN) algorithms will be used to maintain high end-to-end application performance.

These new routers are being designed to operate in the telephony carrier market, where reliability and maintainability are essential considerations. Dual redundant power supplies and dual switching fabrics are just some of the improvements that will prevent system outages due to component failures.

Workstations

The operation of the user workstation is also being changed to enhance its performance and fully integrate it into the network. These enhancements include more-intelligent Network Interface Cards (NICs), operating systems (OS), middleware, and application programming interfaces (APIs). Workstations incorporating these enhancements will operate in the global “computational Grid” environments now being developed.

The intelligent NIC [15] will provide high-performance—instead of simply high-speed—connections to the user workstation. The same low-cost memory subsystems and ASIC controllers used in router design will be found in user workstations. This hardware will provide the underlying mechanisms needed by real-time applications. The details of network packet handling will be shifted to the hardware and away from the main CPU, an arrangement that will provide more deterministic performance and better operation. Feedback from the network switches and routers will be used to modify how packets are injected into the network. Notification may also be passed up to the middleware layer for processing if required.

Coupled with these new NICs are enhancements in the network protocol processing subsystems of the OS. The user application will be allowed to negotiate with the kernel to obtain direct access to the NIC for routine data movement to and from the network. This change will eliminate much of the memory-to-memory copying done in today's systems and, thus, will improve system throughput.

Another change will be adding middleware to the protocol stack to translate the complex network functions into application and user functions. This middleware will coordinate the operation of multiple distributed compute components (i.e., CPUs, storage, and displays), allowing "computational Grid" environments [16] to be developed and deployed. Middleware will give the user workstation the means to control the lower network layers to ensure that real-time application constraints are met. Feedback messages from the lower layers and the middleware of remote components will be processed to modify how packets are generated.

The APIs being developed for real-time and non-real-time applications will allow users to determine and specify the run-time requirements of their applications. These requirements will be passed down to the middleware, where they will be translated into specific network parameters. These parameters will then be used by the NIC to determine when and how to inject packets into the network. Feedback received from remote APIs and applications will be used to modify the behavior of the application to ensure high-performance end-to-end communications.

These advances in workstation design and operation will improve the user environment, making the network an integral and reliable part of the system. In this new environment, users will be able to collaborate easily with distant peers and use network resources that are scattered around the world.

Conclusions

The purpose of Quality of Service (QoS) guarantees is to satisfy user demands for performance. The traditional solution limits user access to the network because core network components are expensive. This limitation takes the form of admission control policies that are established and enforced at every point in the network.

Supporting QoS guarantees in Internet networks by this method will require new and complex admission control policies to handle the dynamic nature of data traffic. These new policies will need to address the lack of an authoritative control center that all the network components can trust. In addition to authentication, the system will also need to support a non-repudiation mechanism that will protect carriers and users from sophisticated attacks. The distributed nature of the Internet control functions makes non-repudiation extremely difficult. When creating such an administratively complex system, both the initial cost of setting it up and the cost of ongoing maintenance need to be considered

A different approach is to exploit recent exponential improvements in network hardwareto create high-capacity, high-performance networks. Important advances include the development of optical Dense Wavelength Division Multiplexers, terabit routers, and intelligent user workstation interfaces. With these new pieces, it will be economically practical to build high-performance networks that can support high QoS guarantees for end-to-end applications without resorting to complex administrative control systems. Instead, such networks will use non-traditional methods of high-speed links and intelligent switches, routers, and user workstations to maintain low congestion probabilities.

These networks will perform better and more reliably than those built on the basis of traditional QoS mechanisms. They will be cheaper to purchase because buying hardware, even the complex new components, is cheaper than developing software to restrict access. They will be easier to maintain because there will be no need to coordinate complex control functions across multiple communicating domains. Finally, they will be easier to upgrade as user or administrative needs change.

In conclusion, we believe that high-performance networks can be used to satisfy application requirements for QoS guarantees in the future Internet. This viewpoint contrasts starkly with the traditional view that QoS guarantees can be met only with complex admission control procedures.

Richard Carlson
Network Research
Argonne National Laboratory

[1] V. Cerf (as told to Bernard Aboba). "How the Internet Came to Be." This article appears in "The Online User's Encyclopedia," by Bernard Aboba. Addison-Wesley, 1993.

[2] K.C. Coffman, A. Odlyzko; The Size and Growth Rate of the internet; First Monday Vol 3 Issue 10

[3] J. Postel. RFC-791 -- Internet Protocol. Sep-01-1981.

[4] J. Nagle. RFC-896 -- Congestion control in IP/TCP internetworks. Jan-06-1984.

[5] W. Stevens. RFC-2001 -- TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms. January 1997

[6] S. Floyd, V. Jacobson; Random early detection gateways for congestion avoidance; Transactions on Networking Vol 1 Number 4; August 1993

[7] R. Braden, D. Clark, S. Shenker. RFC-1633 -- Integrated Services in the Internet Architecture: an Overview. June 1994.

[8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss. RFC-2475 -- An Architecture for Differentiated Services. December 1998.

[9] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, S. Jamin. RFC-2205 -- Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification. September 1997.

[10] K. Nichols, S. Blake, F. Baker and D. Black. RFC-2474 -- Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. December 1998

[11] C. Partridge; "On to Petabit Networks"; Data Communications; http://www.data.com/25years/petabit.html; October 1997

[12] T.W. Chung et-al; Architecture & Engineering Issues for an Optical Internet; http://www.canet2.net/c3/architecture.pdf

[13] D. Al-Salameh et-al; Bell Labs Technical Journal - V3N1 - Jan - Mar 1998; Optical Networking

[14] C. Partridge, et al; A 50-Gp/s Router; IEEE/ACM Transactions on Networking; June 1998

[15] Scheduled Transfer Protocol (ST) ANSI Working Draft T11.1; Project 1245-D; Rev. 2.6; December 1998

[16] I. Foster, C. Kesselman; "The Grid: Blueprint for a New Computing Infrastructure"; Morgan-Kaufmann 1998